*Research Report*

# Confidence Intervals for Proportion Estimates in Complex Samples

Andreas Oranje

Research & Development

July 2006
RR-06-21

# Confidence Intervals for Proportion Estimates in Complex Samples

Andreas Oranje
ETS, Princeton, NJ

July 2006

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html



**Abstract**

Confidence intervals are an important tool to indicate uncertainty of estimates and to give an idea of probable values of an estimate if a different sample from the population was drawn or a different sample of measures was used. Standard symmetric confidence intervals for proportion estimates based on a normal approximation can yield bounds outside the [0,1] scale and poor coverage, because such approximations are generally inappropriate. Many alternative intervals have been proposed to address these issues. This paper discusses a selection of intervals based on scale transformations and continuity corrections and adapts these for use in complex samples. The study expands on the work of Brown, Cai, and DasGupta (2001) and Korn and Graubard (1998) using educational survey designs and complex sample data. Results based on a National Assessment of Educational Progress (NAEP) data resampling study showed that the theoretically appealing Wilson interval yields appropriate coverage with short intervals in most situations.

Key words: Confidence intervals, NAEP, complex samples, variance estimators, jackknife repeated replications

## 1. Introduction

Suppose one has a contingency table with two cells and $n$ observations in the table that are divided between these two cells following a binomial distribution. Furthermore, the proportion of observations in one of the two cells is denoted by $\hat{p}$, and the true proportion is denoted by $p$. Often, symmetric confidence intervals for this proportion are computed following the formula:

$$CI_N = \left(L_N, U_N\right) = p \pm z_{\alpha/2}\sqrt{p\left(1-p\right)/n} \tag{1}$$

where $p$ is approximated by $\hat{p}$, $z_{\alpha/2}$ is the $100\left(1-\alpha/2\right)$ quantile of the standard normal distribution with $\alpha$ usually chosen to be 0.05, and the $\pm$ sign is used to denote the lower bound ($L$) by applying the minus and the upper bound ($U$) by applying the plus. However, (1) is inappropriate, because the proportion scale is bounded, $p \in \left[0,1\right]$, and the scale is discrete, $n\hat{p}, n \in \mathbb{N}^+$. In other words, (1) will yield confidence bounds that often do not exist for a given sample size and confidence level $1-\alpha$. In some cases, this method will yield bounds of the proportion confidence interval that includes points outside $\left[0,1\right]$ (i.e., overshoot). Primarily, the bounded scale would suggest an asymmetric stochastic process.

The discreteness of the scale also results in oscillation of coverage relative to an a priori determined coverage probability $1-\alpha$ for specific $p$ and variable $n$ or for specific $n$ and variable $p$. Coverage is defined as $E\left(p \in \left[L,U\right]\right)$. Despite claims of Brown, Cai, and DasGupta (2001) about the complexity of the oscillation, it is simply a function of the distance between the true proportion and the nearest observable proportions for a given sample size and confidence probability under the binomial distribution (Clopper & Pearson, 1934). As a result, high or low coverage relative to an a priori established nominal level may represent the best possible coverage for a given sample size and not provide substantial information about the success of the confidence interval in providing appropriate coverage. Yet, an extensive query of the binomial distribution (or multinomial distribution in more complex situations) to find appropriate bounds is labor intensive; hence, approximate intervals as often proposed and summarized in Brown et al. have practical virtue.

Several approaches have been developed that generate asymmetric confidence bounds, either from an asymmetric distribution or by centering a symmetric distribution around a value other than the point estimate of the proportion. Some of these approaches apply a continuity correction, whereas other approaches are based on a variance stabilizing function. In some sense, both approaches accomplish the same goal, which is to smooth a discrete function. Only few of the proposed intervals apply a crucial inversion from the observed to the true proportion. First documented by Wilson (1927), the basis for this approach is to state that the probability an observation will occur that falls outside the confidence interval bounds based on the true proportion $p \pm z_{\alpha/2}\sqrt{p(1-p)/n}$ is $\alpha$. This is radically different from the interval in (1) stating that the true value lies outside the observed interval with probability $\alpha$, which is usually between 0 and 1. However, the true proportion lies either inside or outside the interval, and therefore the probability is either 0 or 1.

The purpose of this paper is to apply and compare several methods that aim at providing appropriate confidence intervals to complex sample situations, specifically, stratified, multistage probability samples. A complex sample is defined as a sample of observable, individual units that are distributed over larger, group units, such that the individual units within a group unit are more related to each other than they are to individual units from other group units with respect to variables of interest. Usually, the larger units are called clusters or strata and are selected with a specific probability. Generally, the degrees of freedom of the sampling distribution are sharply reduced, and specific estimators (e.g., jackknife, bootstrap) are needed to establish the appropriate amount of uncertainty in estimates (e.g., Cochran, 1977). These estimators are often aggregates of estimates of uncertainty for each cluster or stratum. Hence, for relatively small samples, substantial variation in a few clusters may dominate the formula or even evoke eccentric behavior, leading to erratic estimates of uncertainty. In sum, finding an appropriate confidence interval for proportions is severely complicated in complex samples because of the reduction in degrees of freedom and a problematic assessment of uncertainty predominantly near the bounds (i.e., 0 and 1) of the scale.

### *1.1 Complex Samples*

Regardless of the type of interval chosen, some kind of provision for complex samples has to be made. In general, a complex sample standard error estimator, $v_{complex}^{\frac{1}{2}}$, can replace a

simple random sample based estimator, $v_{srs}^{\frac{1}{2}}$. This may constitute the replacement of the sample size ($n$) with the effective sample size ($\tilde{n}$) following:

$$\tilde{n} = \frac{n}{d}, d = \frac{v_{complex}}{v_{srs}}$$

The design effect, $d$, is a measure of the grade of complexity of a sample and equal to 1 in a simple random sample. In some cases, predominantly when the number of clusters is small, the design effect is underestimated. A pragmatic solution proposed by Korn and Graubard (1998) is to truncate the effective sample size to the (weighted) sample size to protect against severe underestimation of the design effect. A practical rationale is that in very small groups the frequency of drawing a student may be equal to or less than 1 for each sampling unit (e.g., school). Hence, the sample approaches a simple random sample. However, for typically small but highly clustered groups, this may be unsatisfactory. In general, some degree of overestimation of the effective sample size has to be expected and, subsequently, confidence intervals that are too narrow. If the proportion is equal to 0 or 1, the design effect cannot be estimated at all, and a design effect of 1 can be used in those cases as well.

### *1.2 Confidence Interval Definitions*

Below, the confidence intervals will be listed that were compared in this study, all adjusted for use in complex samples. This list is by no means a complete inventory, but a selection of intervals that have shown reasonable coverage in previous studies. The intervals compared can be divided into exact intervals, approximate intervals, and scale transformations, which also should be considered approximate intervals. Note that student *t* distributions are used instead of normal approximations.

#### *1.2.1 Exact Intervals*

*Wilson.* The idea behind the Wilson interval has been mentioned above and constitutes solving the equation $p \pm t_{df,\alpha/2}\sqrt{p(1-p)/n}$ for $p$ in terms of the observed rate in a quadratic form, $(\hat{p} - p)^2$, yielding the interval:

$$CI_W = \left(L_W, U_W\right) = \frac{\hat{p}n + \frac{1}{2}t_{df}^2}{n + t_{df}^2} \pm \frac{t_{df}\sqrt{n}}{n + t_{df}^2}\sqrt{\hat{p}\left(1-\hat{p}\right) + t_{df}^2 \big/ 4n} \tag{2}$$

One of the problems reported by Brown et al. (2001) with this interval is a sharp downward spike in coverage near the tails of the scale. However, they compared the interval to an a priori set nominal level, which has shown to be misleading (Clopper & Pearson, 1934). In complex samples, a sharp drop also may be encountered due to the probable underestimation of the design effect in those areas. This can be prevented to some extent by setting the minimum design effect at 1 as mentioned earlier.

Brown et al. (2001) proposed a boundary modification. The lower bound is replaced for proportions based on cells with very few observations or the upper bound for proportions based on cells with almost all observations in the sample. The replacement is based on the relationship between a one-sided Poisson and the $\chi^2$ distribution:

$$\begin{aligned} L_{W1} &= \frac{\chi^2_{\nu=2\,\hat{p}\tilde{n},\alpha/2}}{2\tilde{n}} \\ U_{W1} &= \frac{\chi^2_{\nu=2(\hat{p}\tilde{n}+1),1-\alpha/2}}{2\tilde{n}} \end{aligned} \tag{3}$$

Some guidelines are provided to use this replacement for cells with one or two observations if the sample size is less than 50; for cells with one, two, or three observations for sample sizes equal to or larger than 50; and correspondingly for cells with all but one, two, or three observations for the upper bound replacement. This yields the confidence interval $CI_{W1}$. If desired, an estimated design effect can be used to translate these guidelines into effective sample sizes and products of proportion estimates and effective samples sizes, which will be denoted $CI_{W2}$. The bounds in (3)also can be used entirely to compute confidence intervals and are called Breeze intervals $CI_B = \left(L_{W1}, U_{W1}\right)$ (Korn & Graubard, 1998).

*Clopper-Pearson.* The Clopper-Pearson (Clopper & Pearson, 1934) interval is based on the binomial distribution by solving the equation $\sum_{j=x}^{n}\binom{n}{j}p_L^j\left(1-p_L\right)^{n-j} = \frac{\alpha}{2}$ for the lower bound, where *x* and *n* are the summation of binomial variables with common *p*. A complementing

equation is solved for the upper bound, and using Snedecor's *F* results in the solution shown in (4) (e.g., Johnson, Kotz, & Kemp, 1992, p. 130):

$$CI_{CP} = \left(L_{CP}, U_{CP}\right) = \left( \frac{\nu_1 F_{\nu_1\nu_2}\left(\alpha/2\right)}{\nu_2 + \nu_1 F_{\nu_1\nu_2}\left(\alpha/2\right)}, \frac{\nu_3 F_{\nu_3\nu_4}\left(1-\alpha/2\right)}{\nu_4 + \nu_3 F_{\nu_3\nu_4}\left(1-\alpha/2\right)} \right) \tag{4}$$

where $\nu_1 = 2\hat{p}\tilde{n}$, $\nu_2 = 2\left(\tilde{n} - \hat{p}\tilde{n} + 1\right)$, $\nu_3 = 2\left(\hat{p}\tilde{n} + 1\right)$, and $\nu_4 = 2\tilde{n}\left(1 - \hat{p}\right)$ for an *F* distribution with chosen $\alpha$ -level. This procedure is exact in the sense that a similar argument is made as for the Wilson, inversing from an observed proportion interval to a true proportion interval. Furthermore, since the binomial distribution is used, this approach is likely more appropriate for binomial type problems than the Wilson. The only restriction is that an equal tail interval is derived; hence, this interval can be persistently conservative.

### *1.2.2 Approximate Intervals*

*Jeffrey's.* The Jeffrey's prior interval is based on a beta distribution, being the standard prior for binomial distributions. For complex samples, this interval is defined as:

$$CI_J = \left(L_J, U_J\right) = \left( \beta\left(\frac{\alpha}{2}, \nu_1, \nu_2\right), \beta\left(1 - \frac{\alpha}{2}, \nu_1, \nu_2\right) \right) \tag{5}$$

where $\beta$ is a quantile from the (posterior) beta distribution with degrees of freedom $\nu_1 = \hat{p}\tilde{n} + c_1$ and $\nu_2 = \left(1 - \hat{p}\right)\tilde{n} + c_2$ and with chosen level $\alpha$. The prior for the binomial process is $\beta\left(c_1, c_2\right)$, where $c_1 = c_2 = 1/2$ appears to be an appropriate noninformative choice. The following exceptions are applied to ensure that observed proportions equal to 0 or 1 are contained in the confidence interval: $L_J\left(0\right) = 0$ and $U_J\left(n\right) = 1$.

Brown et al. (2001) proposed two consecutive ad hoc boundary modifications. The first modification deals with the situation where all observations are part of the numerator or none are: $U_{J1}\left(0\right) = p_l$ and $L_{J1}\left(n\right) = 1 - p_l$, where $p_l = 1 - \left(\alpha/2\right)^{1/\tilde{n}}$. This adjustment has been shown to be the $100\left(1-\alpha/2\right)\%$ Bayesian prediction interval, based on a uniform prior distribution for $p$ (Johnson et al., 1993, p. 131). The resulting interval will be denoted $CI_{J1}$. The second

adjustment applies to the situation where all but one or one are part of the numerator: $L_{J2}(1)=0$ and $U_{J2}(n-1)=1$. This modification does not bear any particular theoretical underpinnings but has some pragmatic appeal. This interval will be denoted $CI_{J2}$.

*Agresti-Coull.* An interval that applies a continuity correction is the Agresti-Coull (Agresti & Coull, 1998) interval, $CI_{AC}$. This interval is the same as the standard interval in (1), except that $\hat{p}$ is replaced by $\breve{p}=\hat{p}+t_\nu^2/2$ and $n$ by $\breve{n}=\tilde{n}+t_\nu$. These corrections seem somewhat arbitrary, especially since a discrete problem applied to small sample situations likely will not accommodate continuous characteristics easily. Both the Wilson and this interval are centered on the same value. It should be noted that the Agresti-Coull interval can yield overshoot.

### *1.2.3 Scale Transformations*

*Arcsine.* The arcsine is one of the two transformation approaches discussed here. The advantage of a transformation is that complex sample variance estimation methods can be applied to the transformed metric. However, alternative methods are also available using first order Taylor series approximations (see the appendix). Both will be compared. The arcsine transformation, following the function $f_{\hat{p}}=\arcsin\left(\hat{p}^{\frac{1}{2}}\right)$, is a widely used variance stabilizing transformation for the binomial distribution. Straightforward application leads to:

$$CI_A=\sin^2\left(f_{\hat{p}}\pm t_\nu\cdot\sqrt{var\left(f_{\hat{p}}\right)}\right) \tag{6}$$

where $\text{var}\left(f_{\hat{p}}\right)$ can be found by using a complex sample variance estimator such as the bootstrap or the jackknife. A standard arcsine interval based on a first order Taylor series approximation is defined as:

$$CI_{AT}=\sin^2\left(f_{\hat{p}}\pm\tfrac{1}{2}t_\nu\tilde{n}^{-\frac{1}{2}}\right) \tag{7}$$

One important drawback of this approach is the use of a periodic function. If the design effect is poorly estimated, in conceivable cases the lower bound would exceed the upper bound. Also, both these arcsine intervals are nonexistent when the proportion is 0 or 1. Two boundary

modifications have been implemented. The first boundary modification sets the lower bound to 0 if the proportion is 0 and the upper bound to 1 if the proportion is 1. The second modification is similar to the first Jeffrey's boundary modification. These two modifications will be referred to as $CI_{A1}$ and $CI_{AT1}$. Alternatively, a replacement of $\hat{p}$ has been proposed (Anscombe, 1948) for use in the Taylor based logit confidence interval, which in complex samples is defined as $\breve{p} = \left(\hat{p}\tilde{n} + \frac{3}{8}\right)/\left(\tilde{n} + \frac{3}{4}\right)$. This interval is denoted $CI_{AT2}$.

*Logit.* In the logit interval the proportion is transformed using the log odds, and confidence bounds are computed for the logit scale. Reverse transformation of the bounds lead to values in the [0,1] interval. By denoting the logit transformation as $\lambda = \ln\left(\hat{p}/(1-\hat{p})\right)$, a provisional interval in the logit scale based on a first order Taylor series approximation is derived (see the appendix) as :

$$CI_{\lambda T} = \lambda \pm t_{\nu} \frac{\sqrt{var(\hat{p})}}{\hat{p}\hat{q}} \tag{8}$$

where the $\lambda$ subscript denotes that this computation is before reverse transformation. The final confidence interval will be denoted $CI_{LT}$. The transformation stabilizes the variance in the sense that the tails of the scale are widened to allow for a more precise differentiation between very small or very large values on the proportion scale. It should be noted that observed proportions equal to 0 or 1 will result in a nonexisting interval. By inspecting the limits, it can be reasoned that the log of the odds approaches infinity slower than the proportion, resulting in a [0,1] or full-width interval (FWI). The $CI_{LT}$ with application of FWI to specified cases will be denoted by $CI_{LT1}$. These bounds are unsatisfactorily wide; therefore, the first Jeffrey's boundary modification can be applied here as well, yielding $CI_{LT2}$. Furthermore, an alternative estimate for the log odds has been proposed that does not yield a singular result at 0 or unit proportions (Anscombe, 1956), which is $\ln\left(\left(\hat{p}\tilde{n} + \frac{1}{2}\right)/\left(\tilde{n} - \hat{p}\tilde{n} + \frac{1}{2}\right)\right)$. The variance component associated with this interval can be computed as $\left(\left(\tilde{n}+1\right)\left(\tilde{n}+2\right)\right)/\left(\tilde{n}\left(\hat{p}\tilde{n}+1\right)\left(\tilde{n}-\hat{p}\tilde{n}+1\right)\right)$ (proposed by Hadane; see Gart & Zweifel, 1967, estimator $V_3$; see also Cox & Snell, 1989, p. 32, Equation 2.27, and

Santner & Duffy, 1989, p. 32, Equation 2.1.11). This resulting interval after reverse transformation according to the log odds ratio is denoted $CI_{LT3}$.

Alternatively, a complex sample variance estimation method such as the bootstrap or the jackknife can be executed in the transformed metric leading to the interval:

$$CI_{\lambda} = \lambda \pm t_{v}\sqrt{var(\lambda)} \tag{9}$$

followed by the reverse transformation of $CI_L = (L_L, U_L) = \left[\left(1+e^{-L_\lambda}\right)^{-1}, \left(1+e^{-U_\lambda}\right)^{-1}\right]$. Because this interval suffers from similar limit problems as the Taylor based logit similar boundary, modifications can be applied as well to obtain $CI_{L1}$ and $CI_{L2}$.

*Other intervals.* Numerous other approaches to intervals for binomial processes are not part of this study, such as randomized approaches and the likelihood ratio interval. These intervals have not received much attention in the literature but do deserve further study.

### *1.3 Previous Findings*

A few comparative studies have been conducted that are briefly summarized here. Brown et al. (2001) recommended, besides the Agresti-Coull interval, two other intervals: the modified Wilson interval and the modified Jeffrey's interval. Both intervals seem to yield similar and appropriate coverage. Also, they did not recommend the Agresti-Coull interval for small percentages. Newcombe (1998) recommended the Wilson interval using the Clopper-Pearson criterion as a golden standard based on a simulation with simple random samples. Korn and Graubard (1998) studied a limited number of methods for confidence interval calculation around proportions in complex samples. They recommended the Clopper-Pearson interval adjusted for complex samples for use in surveys with relatively few primary sampling units. Specifically, they stated that this interval provides better coverage than a logit transformation based interval. This finding has been replicated by Gray, Haslett, and Kuzmicich (2004), applying the Clopper-Pearson interval to a slightly different type of clustering (i.e., one that approaches a simple random sample rather quickly). It should be noted that Brown et al. have shown the Clopper-Pearson interval has shown to be overly conservative.

### *1.4 Variance Estimation and the National Assessment of Educational Progress*

An assessment that could benefit from asymmetric confidence intervals is the National Assessment of Educational Progress (NAEP). NAEP is the only U.S. assessment that periodically monitors what American students in elementary and secondary schools know and can do in various subjects such as reading and mathematics. NAEP uses a sample of students to estimate average proficiency of subgroups of students in addition to proportions at or above certain cut points of proficiency. These cut points are set by experts under auspices of the National Assessment Governing Board. Also, NAEP provides estimates of subgroup proportions in the population. While uncertainty of estimates in NAEP is exclusively characterized by standard errors, for proportion estimates, overshoot is specifically monitored based on symmetric standard normal intervals, suppressing standard errors if overshoot occurs (Allen, Donoghue, & Schoeps, 2001). Hence, it is pertinent to explore methods to construct asymmetric confidence intervals for NAEP to accurately characterize uncertainty without the need to explain seemingly peculiar results or to suppress results. However, the methodology and findings are widely applicable to other large-scale assessments, such as the Trends in International Mathematics and Science Study (TIMSS) and the National Assessment of Adult Literacy (NAAL).

NAEP draws a stratified multistage probability sample. In the first stage, schools are sampled with probability of selection relative to size for each primary stratum (usually states). In the second stage, a random sample of students is drawn for each sampled school for the age or grade assessed. Schools (and students) are pairwise assigned to replicate strata following systematic ordering based on a limited set of demographics. A jackknife repeated replication (JRR), also referred to as a leave-out-group jackknife, approach is taken to estimate sampling variance by sequentially removing one of the pairs for one replicate stratum at a time. Because only one of the pairs is removed, this approach is more formally known as the JRR-half. The resulting variance estimate accounts for the hierarchical nature of the sample and considers the number of replicate strata to be the maximum degrees of freedom. Student sampling weights are computed as the inverse of the probability of being selected in the sample in combination with several nonresponse adjustments for students and schools.

There are many alternative variance estimators for complex samples (Wolter, 1985). Two popular methods are the bootstrap (e.g., Efron & Tibshirani, 1993) and Taylor series expansion (e.g., Woodruff, 1971). Although the bootstrap is straightforward to implement, large-scale

assessments such as NAEP have been hesitant to implement this approach, predominantly because of computational burden and a relatively small gain in accuracy. Some Taylor series approaches have been studied relatively well in the NAEP context (Hansen & Tepping, 1985; Kovar, 1985; Kovar, Rao, & Wu, 1988). The results generally have shown that variance estimates based on a JRR approach are substantially less biased and more stable compared to stratified Taylor series approximations. Therefore, this study focused exclusively on the JRR approach for estimating variance of proportions or functions of proportions and, subsequently, design effects.

### *1.5 Note on the Sampling Distribution*

In complex samples, the dependencies can reduce severely the degrees of freedom of the distribution of summary statistics, in particular for the cases of interest in this study: small proportions and proportions based on small samples where the expected number of strata can be quite small. However, it is difficult to assess the effective degrees of freedom of the student *t* distribution commonly used for statistical inference. In NAEP, a procedure is used adopted from Satterthwaite (1941) and, subsequently, linearly adjusted for stratum dependencies based on simulation (Johnson & Rust, 1992). The procedure is based on fourth order moments from the JRR procedure and, conversely, has the weakness that underestimation in the standard error also results in underestimation of the degrees of freedom. Moreover, the degrees of freedom of observed proportions equal to 0 cannot be estimated under this procedure, yielding a division by 0, and some boundary correction has to be implemented. For the purpose of this study, a lower bound of 1 to the degrees of freedom has been enforced.

The remainder of this paper is devoted to comparing the methods described above in a NAEP database study. The study compares a wide array of potential intervals, adapting definitions for use in complex samples, by using real data to mimic the complexities of educational surveys as closely as possible.

## 2. Method

In this study, confidence intervals were compared with respect to two quantities: coverage and width. The primary goal was to find methods of computing confidence intervals that provide appropriate coverage, considering conservative and liberal estimates equally problematic. A secondary goal was to select the approach with the shortest confidence interval

widths conditional on appropriate coverage. Conceivably, a small loss in coverage may be traded for a substantial reduction in width. For evaluation, the processes underlying potential oscillation were ignored, as the sample size, discussed below, was substantial, allowing for reasonable precision for most cases of interest.

One of the fundamental challenges this study faced was to simulate a complex sample. This problem was approached by drawing stratified subsamples from a large complex parent sample. The goal behind this procedure is to retain the clustering, sampling design, sampling weights, and structure in addition to several unique sample characteristics that are challenging to account for in a simulation. A disadvantage is that the parent sample is finite; therefore, the stratified samples can be only relatively small to limit dependence, reducing the amount of clustering that can be retained. With respect to the coverage of confidence intervals, one of the effects of this dependence is overestimation of the coverage (i.e., the proportion estimate from the sample drawn will more often fall within the confidence interval than can be expected based on chance). Subsequently, if a relatively large proportion of the parent sample is drawn, coverage will approximate 100%. This effect can be decreased somewhat by sampling with replacement to increase the range of samples, although this may also decrease the clustering. In this study, a without replacement approach was taken for stratified samples that were less than 0.8% of the parent sample size. Another disadvantage of this database approach is little room to control the values of proportions of interest, although it can be argued that the proportions in this study are those of most interest and do span a typical range of proportions that could be associated with difficulties in interval estimation.

### *2.1 Sample*

The 2003 NAEP reading assessment data in Grade 8 was taken as a parent sample (i.e., the population), and all statistics of interest were computed from this population. These were considered the true values. This population contained more than 158,000 students from all 50 states and the District of Columbia. Each state or district was a primary stratum for public schools in addition to several region- and denomination-based primary strata for private schools. Within each stratum, large metropolitan statistical areas were sampled with certainty, while smaller areas were sampled relative to student enrolment. Students in certainty areas were assigned to replicate strata, while in noncertainty areas, schools were assigned. There were 62

replicate strata with two variance units per stratum, which were used for the leave-out-group RRJ method to estimate sampling variance.

For each of 500 repetitions, 100 samples were drawn. In each sample, 10 students were randomly drawn per variance unit, resulting in 1,240 students per sample. This sample size was chosen to minimize the drawbacks of a finite population while analyzing realistic sample sizes. Too large a sample would result in overcoverage, because the sample observations and the population are dependent. For each sample, 95% and 99% confidence intervals were computed, and the true value (i.e., population value) was determined to be inside the interval or not. Subsequently, a coverage percentage (out of 100%) for each repetition was obtained, which was aggregated over repetitions. The jackknife approach in this study, used to compute standard errors and effective sample sizes, is largely the same as that used by NAEP. The replicate stratum and variance unit designations were taken directly from the parent sample.

### *2.2 Statistics of Interest*

The proportion estimates of interest were chosen specifically with three goals in mind: (a) computing both nonextreme and extreme proportions, (b) computing both sample characteristic proportions and proportions based on performance levels, and (c) computing proportions with both unstable numerators and unstable numerators and denominators. The following proportion statistics have been computed:

1. Proportion of students classified by gender and race/ethnicity was computed. Table 1 shows the (weighted) proportions in the parent sample. Two genders and six race/ethnicity categories (White, Black, Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native, Other) combined into 12 proportion estimates.

2. Proportion between performance levels was computed. NAEP's literacy scale was used to determine four proportions of performance levels set at the standard normal distribution quantiles 0.20, 0.60, and 0.95, where the first is the proportion below the first quantile and the fourth is the proportion above the last quantile. The true proportions were calculated from the parent sample.

3. Proportion at each performance level for each race/ethnicity category. These proportions sum to 1 for each race/ethnicity category.

The total number of estimates is 40. All confidence intervals mentioned in the introduction have been computed.

**Table 1**

***Weighted Percentage of Students in a Cross-Tabulation of Gender and Race/Ethnicity Groups in the 2003 Reading Grade 8 NAEP Assessment***

| Race/ethnicity | Gender (%) | | Performance levels (%) | | | |
|---|---|---|---|---|---|---|
| | Male | Female | < I | I | II | > II |
| Total | 50.1 | 49.9 | 19.7 | 38.5 | 37.8 | 4.0 |
| White | 31.6 | 31.0 | 12.4 | 36.8 | 45.5 | 5.2 |
| Black | 8.0 | 8.5 | 35.1 | 42.9 | 21.0 | 1.1 |
| Hispanic | 7.3 | 7.2 | 33.5 | 41.5 | 23.4 | 1.6 |
| Asian/Pacific Islander | 2.2 | 2.2 | 16.2 | 34.8 | 42.5 | 6.5 |
| American Indian/Alaskan Native | 0.6 | 0.6 | 34.6 | 39.6 | 24.0 | 1.7 |
| Other | 0.4 | 0.3 | 16.0 | 41.7 | 38.5 | 3.9 |

*Note.* Performance levels were set at the standard normal distribution quantiles 0.20, 0.60, and 0.95, where the first is the proportion below the first quantile and the fourth is the proportion above the last quantile.

### *2.3 Clustering*

One of the features of a complex sample is a design effect substantially larger than 1. A design effect was computed reflecting the sampling procedure of the simulation study:

$$deff_i^* = \frac{\frac{1}{100}\sum_{j=1}^{100}\left(\hat{p}_{ij} - p\right)^2}{n^{-1}p(1-p)} \qquad (10)$$

where $n$ is the total sample size of 1,240 and $\hat{p}_{ij}$ is the proportion for repetition $i$ and replication $j = 1,2,\ldots,100$. Hence, the numerator is the variation across samples, and the denominator is a simple random sample variance estimate based on the true proportion. Table 2 shows the design effect in the population based on the total sample and using the JRR as numerator, the design effect in the study following (10), and the design effect of the samples as percentage of the

population design effect. From Table 2, it can be concluded that the study was in general reasonably successful in obtaining clustered samples, though not to the same extent as in the population. For all groups, about half to a third of the population clustering was retained.

**Table 2**

***Design Effects for the Population and the Samples and Percentage Clustering Retained in the Samples Compared to the Population***

| Group | Population | Samples | % Clustering retained |
|---|---|---|---|
| Male | | | |
| White | 4.50 | 2.07 | 46% |
| Black | 7.37 | 3.48 | 47% |
| Hispanic | 6.76 | 1.95 | 29% |
| Asian | 9.75 | 3.14 | 32% |
| American Indian | 6.51 | 2.95 | 45% |
| Other | 15.17 | 5.92 | 39% |
| Female | | | |
| White | 3.33 | 1.53 | 46% |
| Black | 7.75 | 3.77 | 49% |
| Hispanic | 7.18 | 2.10 | 29% |
| Asian | 9.17 | 2.77 | 30% |
| American Indian | 5.17 | 2.91 | 56% |
| Other | 4.69 | 2.13 | 45% |
| Performance levels (total sample) | | | |
| Below Level I | 7.47 | 3.05 | 41% |
| At Level I | 2.46 | 1.07 | 43% |
| At Level II | 5.28 | 2.33 | 44% |
| Above Level II | 4.02 | 1.82 | 45% |

## 3. Results

Two statistics will be discussed: coverage and average confidence interval width. Coverage will be discussed first, followed by width.

### *3.1 Coverage*

Using the 40 statistics of interest discussed in the previous sections, the average, minimum, and maximum coverage as well as deviations from the nominal level were computed. In Table 3 the coverage results are summarized. The first three results columns display results for 99% confidence intervals. These results reveal that, besides poor average coverage, the arcsine-based intervals showed tremendous variation in coverage. Also, the disadvantages of the standard confidence interval were characterized; the Breeze interval performed slightly worse compared to most other intervals, with intervals that could cover as little as 90%, when 99% was intended.

The last six columns in Table 3 are similar to the first few columns, except that the normal, arcsine, and Breeze confidence intervals are removed, limiting the results to a subset of proportions: those that had an expected sample size less than 3% of the total sample in the numerator. Furthermore, in addition to 99% confidence intervals, 95% confidence intervals are displayed. The results for larger sample sizes were highly satisfactory and remarkably similar across all methods, including the normal and arcsine. The results showed that several intervals did not provide adequate coverage in addition to the arcsine, normal, and Breeze, when considering relatively small samples. These were mainly the Jeffrey's intervals, showing relatively large undercoverage. Also, the first three logit intervals—CI(L), CI(L1), and CI(L2)—applying a complex sample variance estimation technique directly, seemed to result in large undercoverage.

Some intervals provided more satisfactory average coverage and were remarkably close to each other: the Wilson intervals and the Taylor series expansion based logit intervals, especially under Anscombe's adjustment. It is interesting to note that the boundary adjustments in both sets of confidence intervals had little impact on the coverage. Also, the Agresti-Coull and the Clopper-Pearson confidence intervals appeared to provide satisfactory coverage on average, although the distribution of coverage in both these cases seemed much wider. For example, the maximum coverage for both intervals was at or close to 100%. However, this was also true to some extent for the Taylor series expansion based logit and the Wilson confidence intervals. On a related note, the maximum coverage of 100% of the logit intervals LT and LT1 were inspected to make sure that for some subgroups all replications were not FWI. This was not the case.

**Table 3**

***Average, Minimum, and Maximum Coverage Deviation From Nominal Level on a Scale of [0,1] Across 40 and 20 Subgroups for 99% and 95% Confidence Intervals***

| Confidence interval | All subgroups ($n$ = 40) | | | Subgroups with expected sample size < 3% of total sample ($n$ = 20) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 99% | | | 99% | | | 95% | | |
| | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| Normal | -0.069 | -0.647 | 0.000 | | | | | | |
| W | 0.001 | -0.013 | 0.009 | 0.002 | -0.013 | 0.009 | 0.010 | -0.036 | 0.038 |
| W1[a] | 0.001 | -0.013 | 0.009 | 0.002 | -0.013 | 0.009 | 0.013 | -0.036 | 0.044 |
| W2[b] | 0.001 | -0.013 | 0.009 | 0.002 | -0.013 | 0.009 | 0.013 | -0.036 | 0.044 |
| B | -0.012 | -0.078 | 0.009 | | | | | | |
| J | -0.011 | -0.047 | 0.006 | -0.016 | -0.047 | 0.006 | -0.042 | -0.098 | 0.028 |
| J1[c] | -0.010 | -0.047 | 0.007 | -0.015 | -0.047 | 0.007 | -0.034 | -0.079 | 0.028 |
| J2[d] | -0.010 | -0.047 | 0.007 | -0.015 | -0.047 | 0.007 | -0.033 | -0.079 | 0.035 |
| CP | -0.004 | -0.038 | 0.009 | -0.005 | -0.038 | 0.009 | 0.002 | -0.048 | 0.044 |
| A | -0.193 | -0.656 | -0.001 | | | | | | |
| A1[e] | -0.178 | -0.650 | -0.001 | | | | | | |
| AT[f] | -0.134 | -0.979 | 0.001 | | | | | | |
| AT1[g] | -0.090 | -0.499 | 0.001 | | | | | | |
| AT2[h] | -0.119 | -0.923 | 0.001 | | | | | | |

*(Table continues)*

Table 3 (continued)

| Confidence interval | All subgroups ($n$ = 40) | | | Subgroups with expected sample size < 3% of total sample ($n$ = 20) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 99% | | | 99% | | | 95% | | |
| | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max |
| L | -0.004 | -0.027 | 0.004 | -0.008 | -0.027 | 0.004 | -0.027 | -0.054 | -0.003 |
| L1[i] | -0.003 | -0.027 | 0.006 | -0.007 | -0.027 | 0.006 | -0.021 | -0.053 | 0.030 |
| L2[j] | -0.004 | -0.027 | 0.006 | -0.008 | -0.027 | 0.006 | -0.021 | -0.053 | 0.030 |
| LT[k] | 0.003 | -0.007 | 0.010 | 0.005 | -0.007 | 0.010 | 0.019 | -0.031 | 0.043 |
| LT1[l] | 0.003 | -0.007 | 0.010 | 0.006 | -0.007 | 0.010 | 0.022 | -0.031 | 0.043 |
| LT2[m] | 0.003 | -0.007 | 0.009 | 0.005 | -0.007 | 0.009 | 0.022 | -0.031 | 0.043 |
| LT3[n] | 0.002 | -0.008 | 0.009 | 0.004 | -0.008 | 0.009 | 0.011 | -0.035 | 0.031 |
| AC | 0.002 | -0.010 | 0.010 | 0.004 | -0.010 | 0.010 | 0.019 | -0.029 | 0.047 |

*Note.* A = Arcsine; AC = Agresti-Coull; B = Breeze; CP = Clopper-Pearson; J = Jeffrey's; L = logit; LT = logit based on a Taylor series expansion; W = Wilson.

[a] Adjusted Wilson based on counts. [b] Adjusted Wilson based on effective sample size. [c] Adjusted Jeffrey's for U(0) and L(n). [d] Adjusted Jeffrey's for U(0), L(n), L(1) and U(n-1). [e] Adjusted arcsine for L(0), U(n), U(0), L(n). [f] Arcsine based on a Taylor series expansion. [g] Arcsine based on a Taylor series expansion and adjusted for L(0), U(n), U(0), L(n). [h] Arcsine using Anscombe's adjustment (Taylor series expansion). [i] Adjusted logit adding full-width interval (FWI). [j] Adjusted logit adding FWI and adjustments for U(0) and L(n). [k] Logit based on a Taylor series expansion. [l] Logit based on a Taylor series expansion and adding FWI. [m] Logit based on a Taylor series expansion and adding FWI and adjustments for U(0) and L(n). [n] Logit using Anscombe's adjustment (Taylor series expansion).

For illustration purposes, the Anscombe adjusted logit and the Wilson were compared directly. Figure 1 compares all groups of interest for 99% intervals. This figure reveals that the Wilson and logit intervals were relatively close to each other and to nominal levels, except for the performance levels of Asian American students and students classified as Other, where the Wilson interval was substantially closer to the nominal level.

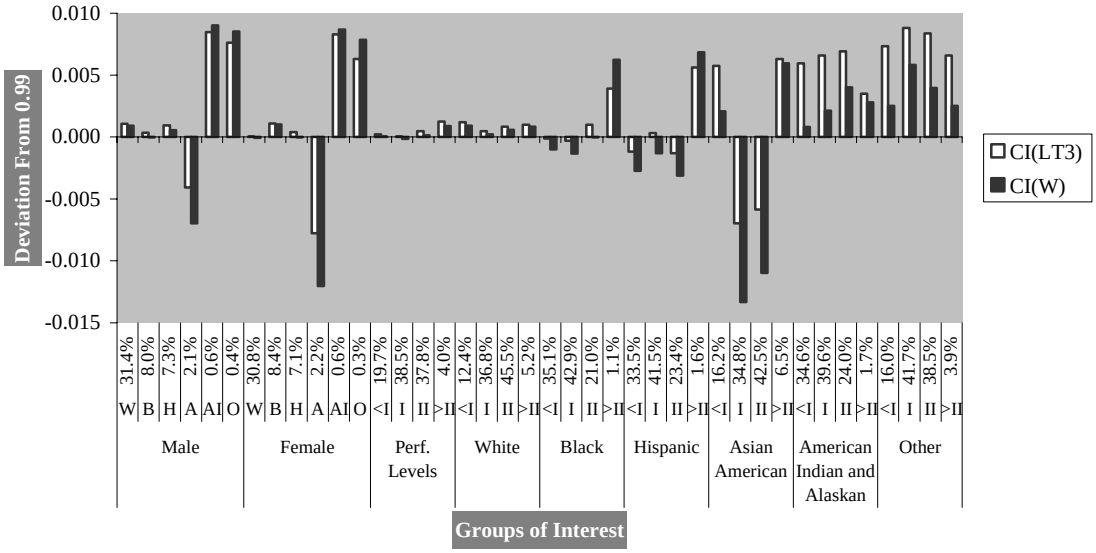


***Figure 1.*** **Deviation from nominal coverage of the Wilson and logit (LT3) 99% confidence intervals for 40 estimates of interest.**

### *3.2 Width*

The second statistic used for evaluation was the width of the intervals. To summarize the results, for each group of interest the width of each interval method was divided by each other interval method. Subsequently, these ratios were averaged across groups and averaged across ratios that shared the same method in the numerator. In formula, if there are $k = 1,2..,K$ proportions estimated, and there are $m = 1,2,\ldots,M$ methods, then the average width ratio for method $j$ is the following:

$$\overline{r}_j = \frac{1}{K(M-1)} \sum_{m \neq j} \sum_{k=1}^{K} \frac{(U_{kj} - L_{kj})}{(U_{km} - L_{km})}$$

These averages are shown in Table 4 across all groups, across the 20 groups that were based on less than 3% of the sample (numerator), and across the 20 groups that were based on more than 3%. The basic interpretation was that a smaller interval would have a lower ratio; in this case, the Jeffrey's intervals (columns J, J1, and J2) had the smallest average ratios. This was to be expected, because these intervals also had the most liberal coverage, likely resulting in intervals that were too short. Between the Wilson and the logit it is clear that the Wilson has on average smaller interval bounds. Among the Wilson intervals, the unmodified Wilson (W) had the smallest interval; among the logit, the LT3 had the smallest intervals. Also, for large groups essentially all interval widths, except possibly the Jeffrey's, were similar, leading to ratios close to 1. The ratio of the Wilson over the logit was 0.91 for all observations and 0.82 for the small sample subset.

## 4. Summary and Conclusion

In this study, several approaches to computing confidence intervals for proportion estimates in complex samples have been compared. Three related issues, regardless of the complexity of the sample, usually surface: (a) the categorical nature of the scale, (b) the bounds of the scale (overshoot), and (c) the accuracy of variance estimation at the extremes of the scale (degeneration). The result of these phenomena is oscillation of the coverage along the proportion scale relative to a priori determined nominal levels, even for relatively large sample sizes and moderate proportions. Several methods have been proposed in the literature to obtain a priori nominal coverage. Some of these methods entail the application of a continuity correction based on some continuous (normal) distribution derived interval, whereas other proposed solutions include transformations of the metric of the proportion scale. Wilson (1927) and Clopper and Pearson (1934), formulating the problem in a radically different fashion, have provided a conceptually appealing approach, resulting in exact as opposed to approximate confidence intervals. Whereas each approach addresses in part different issues or the same issues in different ways, the resulting coverage and bounded intervals make an interesting comparison.

It is important to put the notion of a priori nominal coverage into perspective. With a small sample size of, say, 5 observations, it is obvious that coverage cannot be nominal. In fact, for most choices of sample size, true proportion, and alpha, nominal coverage is not possible and the best coverage can be substantially less or more than the a priori nominal coverage. As discussed in the introduction, this is the process that generates the oscillation effect and is a

**Table 4**

***Average Width Ratios for 15 Confidence Interval Methods, for all Groups, Those Groups That Represent Less Than 3% of the Population, and Those That Represent More Than 3% of the Population***

| Interval | W | W1[a] | W2[b] | J | J1[c] | J2[d] | CP | L | L1[e] | L2[f] | LT | LT1[g] | LT2[h] | LT3[i] | AC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 99% | | | | | | | | | | | | | | | |
| All | 1.08 | 1.09 | 1.09 | 0.80 | 0.80 | 0.80 | 0.85 | 1.53 | 1.57 | 1.49 | 1.44 | 1.47 | 1.39 | 1.29 | 1.16 |
| < 3% | 1.16 | 1.16 | 1.16 | 0.65 | 0.67 | 0.67 | 0.73 | 2.03 | 2.11 | 1.95 | 1.86 | 1.92 | 1.76 | 1.55 | 1.29 |
| > 3% | 1.01 | 1.01 | 1.01 | 0.94 | 0.94 | 0.94 | 0.97 | 1.03 | 1.03 | 1.03 | 1.03 | 1.03 | 1.03 | 1.02 | 1.02 |
| 95% | | | | | | | | | | | | | | | |
| All | 1.03 | 1.03 | 1.03 | 0.83 | 0.84 | 0.84 | 0.91 | 1.39 | 1.47 | 1.36 | 1.25 | 1.32 | 1.21 | 1.24 | 1.08 |
| < 3% | 1.05 | 1.06 | 1.06 | 0.70 | 0.72 | 0.72 | 0.81 | 1.77 | 1.93 | 1.70 | 1.49 | 1.63 | 1.40 | 1.47 | 1.16 |
| > 3% | 1.01 | 1.01 | 1.01 | 0.96 | 0.96 | 0.96 | 1.00 | 1.02 | 1.02 | 1.02 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 |

*Note.* AC = Agresti-Coull; CP = Clopper-Pearson; J = Jeffrey's; L = logit; LT = logit based on a Taylor series expansion; W = Wilson.

[a] Adjusted Wilson based on counts. [b] Adjusted Wilson based on effective sample size. [c] Adjusted Jeffrey's for U(0) and L(n). [d] Adjusted Jeffrey's for U(0), L(n), L(1) and U(n-1). [e] Adjusted logit adding full-width interval (FWI). [f] Adjusted logit adding FWI and adjustments for U(0) and L(n). [g] Logit based on a Taylor series expansion and adding FWI. [h] Logit based on a Taylor series expansion and adding FWI and adjustments for U(0) and L(n). [i] Logit using Anscombe's adjustment (Taylor series expansion).

function of the distance between the true proportion and the closest observable proportions for a given sample size. The only confidence intervals that take this into account are the Wilson and the Clopper-Pearson intervals. Hence, applying corrections for certain boundary cases (e.g., one or two observations in the cell of interest), as Brown et al. (2001) proposed frequently, seems misguided, because a priori nominal coverage is not necessarily appropriate as a reference. In some sense, it seems surprising that so much effort has been put into developing alternative confidence intervals, when the problem seems to have been largely solved in 1927. The findings of this study show that these adjustments essentially do more harm than good to the Wilson interval. Hence, it is not surprising that the Wilson interval provides the best results in terms of coverage and width in this study.

In formulating a recommendation on the use of confidence intervals for proportion estimates in complex samples, no single method can be advised without further consideration of specific requirements. Also, the specific properties of the chosen complex sample variance estimation technique have to be taken into consideration. There is no doubt that the JRR approach has influenced the outcome of this study substantially. Variance estimates based on the JRR method tend to become increasingly unreliable if relatively few primary sampling units are used (Burke & Rust, 1995), and thus the degrees of freedom of the proportion distribution is small. This is especially cumbersome in highly clustered populations, such as American Indians in the studied data, who are distributed among a small set of primary sampling units. Furthermore, this problem is carried forward in the degrees of freedom estimate, using a Satterthwaite estimator (Johnson & Rust, 1992; Satterthwaite, 1941) to determine the *t*-distribution quantiles. Subsequently, the question arises: To what extent is the sample appropriate for estimating characteristics of certain subgroups? Nevertheless, the results from this study in combination with the theoretical background of each of the methods compared suggest that the Wilson approach is most appropriate.

The key issue addressed in this paper is how these approaches can be adapted for use in complex samples (e.g., clustered multistage probability samples) and how these methods compare. The study was based on the methodology used in several large-scale assessments such as the NAEP, the NAAL, the TIMSS, and the Progress in International Reading Literacy Study. In most of these studies, students are clustered within schools, and schools are clustered within relatively homogenous geographic areas with respect to median income, racial/ethnic

distribution, language, and so on. Taking NAEP as the lead example, variance estimation in this study was conducted using a JRR approach.

The results showed that for most moderate proportions (more than 5% or less than 95%) and reasonably large sample sizes (about 50 clustered observations or more) all studied confidence intervals except the standard and arcsine approaches provide nominal coverage. Confidence intervals that also provide nominal coverage for small proportions and small sample sizes are the Wilson, Clopper-Pearson, and Agresti-Coull intervals as well as the Taylor series based logit interval with Anscombe's continuity correction. Among these, the Wilson approach is generally closest to nominal coverage. Also, the width of the Wilson confidence interval is on average the smallest. The Clopper-Pearson interval is relatively conservative due to the equal tail restriction, and the Agresti-Coull interval does not solve satisfactorily the problem of overshoot. Finally, the logit approach shows substantially wider confidence intervals than the Wilson approach.

## References


Agresti, A, & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician, 52,* 119-126.

Allen, N., Donoghue, J. R., & Schoeps, T. (2001). *The NAEP 1998 technical report.* Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Education Statistics.

Anscombe, F. J. (1948). The transformation of Poisson, binomial, and negative binomial data. *Biometrika, 35,* 246-254.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101-133.

Burke, J., & Rust, K. (1995). On the performance of jackknife variance estimation for systematic samples with small numbers of primary sampling units. *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 321-327). Alexandria, VA: American Statistical Association.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26,* 404-413.

Cochran, W. G. (1977). *Sampling techniques*. New York: John Wiley & Sons.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Hansen, M. H., & Tepping, B. (1985). *Note to professor N. J. K. Rao.* Unpublished manuscript.

Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika, 54*(1&2), 181-187.

Gray, A., Haslett, S., & Kuzmicich, G. (2004). Confidence intervals for proportions estimated from complex sample designs. *Journal of Official Statistics, 20*(4), 705-723.

Johnson, E. G., & Rust, K. F. (1992). Effective degrees of freedom for variance estimates from a complex sample survey. *Proceedings of the Survey Research Methods Section, American Statistical Association* (pp. 863-866). Alexandria, VA: American Statistical Association.

Johnson, N. L., Kotz, S., & Kemp A. W. (1993). *Univariate discrete distributions* (2nd ed.). New York: John Wiley & Sons.

Korn, E. L., & Graubard, B. I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology, 24*(2), 193-201.

Kovar, J. (1985). *Variance estimation of nonlinear statistics in stratified samples* (Working paper No. BSMD 85-052E). Toronto, Ontario, Canada: Statistics Canada.

Kovar, J. G., Rao, J. N. K., & Wu, C. F. J. (1988). Bootstrap and other methods to measure error in survey estimates. *The Canadian Journal of Statistics, 16,* 25-45.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17,* 857-872.

Santner, T. J., & Duffy, D. E. (1989). *The statistical analysis of discrete data.* New York: Springer.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309-316.

Wilson, E. B. (1927). Probably inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22,* 209-212.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association, 66*(334), 411-414.

## Appendix

## Derivation of Taylor Series Based Intervals

The logit transformation confidence interval formula in logit metric is the following:

$$CI_{\text{logit}} = \ln\left(\frac{\hat{p}}{\hat{q}}\right) \pm z\frac{\sqrt{Var(\hat{p})}}{\hat{p}\hat{q}}$$

where $\hat{p} = (1-\hat{q})$ is the estimated proportion in the sample of the population with proportion value $\pi$ and $z$ is the 97.5 or 99.5 percentile of the standard normal distribution or an appropriate Student *t* distribution. The formula in (8) is based on a first order Taylor series approximation. If $f(x)$ is the logit transformation for *x*, then

$$\ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{\hat{p}}{\hat{q}}\right) + \frac{f'(\hat{p})}{1!}\cdot(\pi-\hat{p}) = \ln\left(\frac{\hat{p}}{\hat{q}}\right) + \frac{1}{\hat{p}\hat{q}}\cdot(\pi-\hat{p})$$

because $f'(\hat{p}) = \dfrac{\partial \ln\left(\frac{\hat{p}}{\hat{q}}\right)}{\partial \hat{p}} = \dfrac{1}{\hat{p}\hat{q}}$

$$\Leftrightarrow \ln\left(\frac{\pi}{1-\pi}\right) - \ln\left(\frac{\hat{p}}{\hat{q}}\right) = \frac{1}{\hat{p}\hat{q}}\cdot(\pi-\hat{p})$$

$$\Leftrightarrow \left(\ln\left(\frac{\pi}{1-\pi}\right) - \ln\left(\frac{\hat{p}}{\hat{q}}\right)\right)^2 = \frac{1}{(\hat{p}\hat{q})^2}\cdot(\pi-\hat{p})^2$$

$$\Leftrightarrow E\left[\left(\ln\left(\frac{\pi}{1-\pi}\right) - \ln\left(\frac{\hat{p}}{\hat{q}}\right)\right)^2\right] = \frac{1}{(\hat{p}\hat{q})^2}\cdot E\left[(\pi-\hat{p})^2\right]$$

$$\Leftrightarrow Var\left[\ln\left(\frac{\hat{p}}{\hat{q}}\right)\right] = \frac{1}{(\hat{p}\hat{q})^2}\cdot Var[\hat{p}]$$

Therefore, the standard error is $\frac{\sqrt{Var(\hat{p})}}{\hat{p}\hat{q}}$.

The arcsine transformation confidence interval in arcsine metric is the following:

$$CI_{\arcsin} = \arcsin\left(\sqrt{\hat{p}}\right) \pm z\frac{1}{2\sqrt{\tilde{n}}}$$

where $\tilde{n} = \frac{n}{(deff)} = \frac{\hat{p}\hat{q}}{Var_{complex}(\hat{p})}$ and (*deff*) is the design effect. If $f(x)$ is the arcsine of the square root of *x* transformation, then

$$\begin{aligned}\arcsin\left(\sqrt{\pi}\right) &= \arcsin\left(\sqrt{\hat{p}}\right) + \frac{f'(\hat{p})}{1!}\cdot(\pi - \hat{p}) \\ &= \arcsin\left(\sqrt{\hat{p}}\right) + \frac{1}{2\sqrt{\hat{p}\hat{q}}}\cdot(\pi - \hat{p})\end{aligned}$$

because $\frac{\partial \arcsin(x)}{\partial x} = \frac{\partial \sin^{-1} x}{\partial x} = \frac{1}{\sqrt{1-x^2}}$, and therefore $\frac{\partial \arcsin\left(\sqrt{\hat{p}}\right)}{\partial \hat{p}} = \frac{1}{\sqrt{\hat{q}}}\cdot\frac{1}{2\sqrt{\hat{p}}}$.

$$\Leftrightarrow \arcsin\left(\sqrt{\pi}\right) - \arcsin\left(\sqrt{\hat{p}}\right) = \frac{1}{2\sqrt{\hat{p}\hat{q}}}\cdot(\pi - \hat{p})$$

$$\Leftrightarrow Var\left[\arcsin\left(\sqrt{\hat{p}}\right)\right] = \frac{1}{4\hat{p}\hat{q}}\cdot Var[\hat{p}] = \frac{1}{4\tilde{n}}$$

Therefore, the standard error is $\frac{1}{2\sqrt{\tilde{n}}}$.